of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Loftus, E. F., Loftus, E., & Ketcham, K. (1992). Witness for the defense: The accused, the eyewitness, and the expert who puts memory on trial. New York, NY: St. Martin's Griffin.

Mead, M. (1935). Sex and temperament in three primitive societies. New York, NY: Morrow.

Mroczek, D. K., & Spiro III, A. (2005). Change in life satisfaction during adulthood: findings from the veterans affairs normative aging study. *Journal of Personality and Social Psychology, 88*(1), 189.

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3*(102).

Oakley, A. (1972). *Sex, Gender, and society.* London: Temple Smith.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation, 7*(1).

Osborne, J. W. (2003). Effect Sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, and Evaluation, 8*(99).

Osborne, J. W. (2008). Is disattenuation of effects a best practice? In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 239–245). Thousand Oaks, CA: Sage.

Osborne, J. W. (2012). Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data. Thousand Oaks, CA: Sage.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation, 8*(2).

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction. Fort Worth, TX: Harcourt Brace College.

Quirk, K. J., Keith, T. Z., & Quirk, J. T. (2001). Employment during high school and student achievement: Longitudinal analysis of national data. *The Journal of Educational Research, 95*(1), 4–10.

Rescorla, L., & Rosenthal, A. S. (2004). Growth in standardized ability and achievement test scores from 3rd to 10th grade. *Journal of Educational Psychology, 96*(1), 85.

Sullivan, S. E., & Bhagat, R. S. (1992). Organizational stress, job satisfaction and job performance: Where do we go from here? *Journal of Management, 18*(2), 353–374.

Teigen, K. H. (1994). Yerkes-Dodson: A law for all seasons. *Theory & Psychology, 4*(4), 525–547.

Yegiyan, N. S., & Lang, A. (2010). Processing central and peripheral detail: How content arousal and emotional tone influence encoding. *Media Psychology, 13*(1), 77–99.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(5), 459–482.

2

HOW DOES LOGISTIC REGRESSION HANDLE A BINARY DEPENDENT VARIABLE?

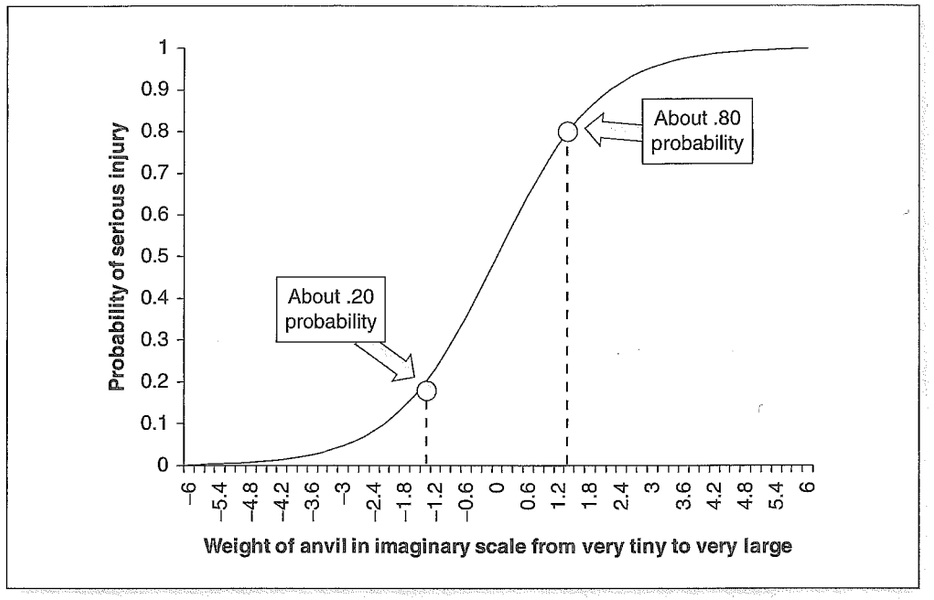
In what way is having an anvil dropped on your head like a logistic curve? What predicts whether students will drop out of school prior to graduating high school? Put a more positive way, what factors influence whether students are retained through high school graduation?

Imagine we had an infinite variety of anvils that we could drop on your head.[1] If we start with an infinitely tiny anvil, and gradually increase the size (and weight) of the anvils that get dropped on your head, for a long while you would not notice or care that extremely tiny things were being dropped on you. Then, suddenly, the anvils would become annoying, then painful, then pretty damaging. And just as quickly, they would become so damaging that increasing the size of the anvil really didn't matter. The result would be terminal headache beyond a certain point (unless you are Wile E. Coyote, in which case you see stars, a comically large lump grows out of your head, and then you come back for another adventure) (see Figure 2.1).

From reading the first chapter, you know that both ordinary least squares (OLS) regression and logistic regression share some important

[1]Something that seems to happen frequently to Wile E. Coyote in Looney Tunes cartoons, as well as to Larry the Cucumber in one of the Osborne family's favorite episodes of *Veggie Tales*.

Figure 2.1 Imaginary Relationship Between Anvil Size and Probability of Serious Injury



commonalities. Most relevant to this chapter, both assume a *linear* relationship between a predictor variable (or set of predictor variables) and a dependent variable. In OLS regression, it is relatively simple to imagine how that relationship is estimated. But how does a binary 0/1 or yes/no variable get handled appropriately in logistic regression? In this chapter, we will look "under the hood" so to speak, examining how counts and probabilities and odds and odds ratios and logits come to be.

The simple example we will use through this chapter is from the National Education Longitudinal Study of 1988 (NELS88) from the National Center for Educational Statistics (NCES; http://nces.ed.gov/surveys/nels88/), a survey of students in eighth grade in the United States in 1988. These students were followed for many years on thousands of variables, similar to other studies from NCES. In particular, we will predict DROPOUT before completing 12th grade (1 = yes, 0 = no)[2] from the variable

[2]For those of you who are interested, I considered students who dropped out and returned as dropouts as well.

POOR (1 = the student falls below the average family income or 0 = the student falls above the average family income).[3]

PROBABILITIES, CONDITIONAL ♦ PROBABILITIES, AND ODDS

There will be a bit of hand calculation in this chapter. This is becoming rarer in statistics texts, with the modern amenities of easy access to powerful statistical computing packages. But in this case, I find it instructive and important to understanding what is happening in logistic regression. Knowing how to manipulate probabilities and odds and odds ratios and logits may help you be more successful (and accurate) when presenting your results.

Table 2.1 Cross-Tabulation of Family Income and Dropout

		Dropout		Total	Conditional probabilities	Odds	Odds ratio (change in odds)
		0 (no)	1 (yes)				
Poor	Yes (1)	7312	1244	8556	0.145	0.170	5.67
	No (0)	7821	233	8054	0.029	0.030	
Total		15133	1477	16610	0.089		

Data Source: NELS88, NCES, U.S. Department of Education (http://nces.ed.gov/surveys/nels88/).

Let's start with a simple example of student family affluence and dropout before completion of high school. Looking at Table 2.1, we can see the number of students from poor or not poor households who dropped out and who were graduated, as well as the overall numbers. It is then relatively simple to start thinking about probabilities. The probability of an event is calculated as the frequency of the event divided by the total observations (in this case, 1,477 dropouts out of 16,610 total students).

[3]I will remind you that I am using this public data for demonstration purposes only. I intentionally did not weight the data or do any of the methodologically important steps necessary to use these data for drawing substantive conclusions. For more on the importance of weighting complex samples such as this, I will refer you to my paper on the topic: http://pareonline.net/pdf/v16n12.pdf (Osborne, 2011).

Probability of dropout ($P_{dropout}$) = number dropouts / total students

$$P_{dropout} = 1477 / 16610$$

$$P_{dropout} = 0.0889$$

When there are two categories (as with this dropout/graduated variable), the probability of a student falling into the "graduated" category is $(1 - P_{dropout}) =$

Probability of graduated $(1 - P_{dropout})$ = number graduated / total students

or $1 - P_{dropout}$

$$1 - P_{dropout} = 15133 / 16610 \textit{ or } 1 - 0.0889$$

$$1 - P_{dropout} = 0.9111$$

According to these data, the overall probability that any student in the eighth grade will drop out is 0.0889, and the probability that any random student will complete a high school education is 0.9111. Knowing nothing else about any student, the best we could do is guess that each student has about a 9% chance of dropping out before graduating high school and a 91% probability of staying in through completion. Of course, we know that the probabilities are not the same for all students. For example, looking at Table 2.1, it is relatively obvious that family income has an effect on the probabilities of dropping out. Thus, we can calculate *conditional probabilities,* a fancy phrase that simply means the probability of something happening within a condition. Thus, for example, we can calculate the conditional probability of dropout for students coming from "poor" households, or households with below-average income (1,244 students in this group dropped out from a total of 8,556), yielding a probability of 0.145. Likewise, we can calculate the conditional probability for those students coming from households with above-average income (233 students in this group dropped out from a total of 8,054), yielding a probability of 0.029. In other words, by knowing one piece of information about a student's background, we know a lot more about the probability of dropout. Students coming from below-average income households are much more likely to drop out than students coming from above-average income households.

In fact, those of you with a background in OLS regression might find it interesting to note that when you have dichotomous variables in OLS regression, with both variables coded 0 and 1, the conditional probabilities of dropout is the predicted variable. Putting the exact same data into an OLS regression analysis produces the following results:

Table 2.2 OLS Regression Results of the Family Income and Dropout Data

	Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
	B	Std. Error	Beta		
(Constant)	.029	.003		9.318	.000
poor	.116	.004	.204	26.923	.000

Data Source: NELS88, NCES, U.S. Department of Education (http://nces.ed.gov/surveys/nels88/).

And it allows the following prediction equation:

Conditional probability of dropout = 0.029 + 0.116 (Poor)

As you can see from Table 2.1, when the independent variable (IV) is 0 (not poor), the conditional probability is 0.029, which matches our calculated conditional probability in Table 2.2. Likewise, when POOR = 1, the predicted probability is 0.029 + 0.116, or 0.145, the predicted probability of the other group.

As we mentioned in the previous chapter, OLS regression has been used for this type of analysis for many years, although the assumptions are difficult to meet, and the predicted probabilities can become impossible when the IV is continuous (i.e., below 0 or above 1.0). To demonstrate this, I substitute the continuous variable (family socioeconomic status [SES], which is *z*-scored) into the OLS regression equation, which gives us the following regression equation:

Conditional probability = 0.089—0.075(SES)

Note that because SES is *z*-scored, the average family income is 0.00, which gives us the conditional probability for a student dropping out when family income is average—exactly the 0.089 we calculated for the overall probability of dropout for the entire sample.

So we can calculate the conditional probability of dropout for a student who is one standard deviation below the mean by substituting −1 into the equation (above) for SES. This yields a conditional probability

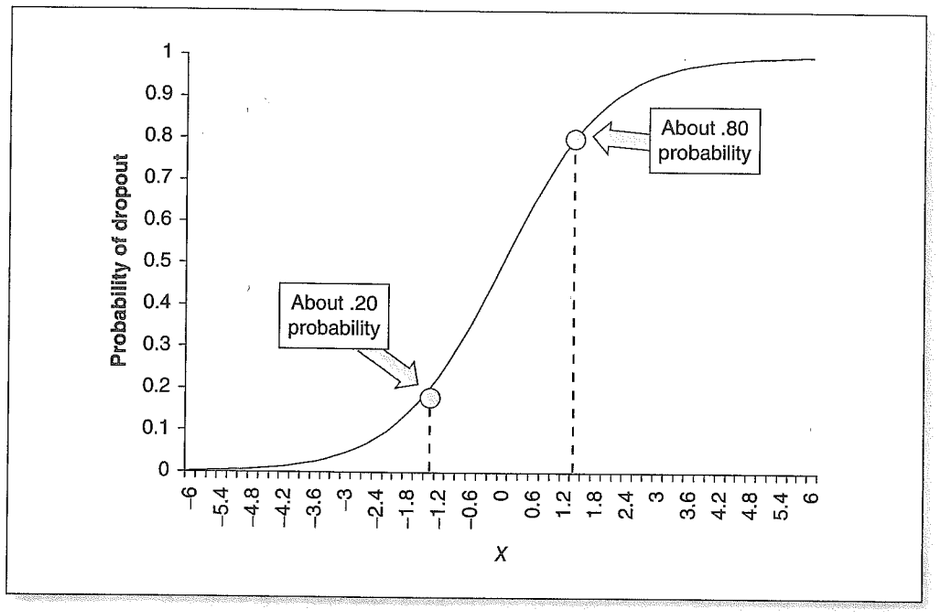
of 0.164, meaning that the probability of dropout for students who are coming from very poor families is much higher than average, and likewise, substituting 1 into the equation yields a conditional probability of 0.014. So far, so good. As expected from our cross-tabulation, poor students have higher probabilities of dropout than more affluent students. The problems come as we begin predicting more extreme cases. The conditional probability for a student 2 *SD* below the mean is 0.239, but for 2 *SD* above the mean is −0.061 (an impossible probability—0.00 is the lowest possible probability). The issues get more extreme as we move toward the extremes of the distribution—the conditional probability for a student coming from a very wealthy family 3 *SD* above the mean is −0.136. Functionally, this means that the probabilities are *really small,* which matches our real-world data and intuition. But the predicted probabilities still violate mathematical rules and assumptions of OLS regression.

A Brief Thought Experiment on the Logistic Curve

From these OLS regression data and common sense, we can see something that is usually presented in discussions of logistic regression but not delved into deeply: the logistic curve. Imagine a theoretical world where poverty was very strongly related to the probability that a student would drop out. As you move from average income to higher and higher poverty, the probabilities move toward 1.0, but at some point, increased poverty doesn't substantially increase the probability of dropout. There may be a threshold after which it really doesn't matter much. Conversely, as you move downward from average toward extremely low poverty, you will quickly asymptote toward 0, and after a certain level of affluence (or lack of poverty), moving toward lower levels of poverty doesn't substantially decrease the probability of dropout. This theoretical relationship is presented below in Figure 2.2. As you can see, there is a relatively narrow window of poverty where changing makes a large difference, and outside that window, the probabilities don't change a great deal. In this fictitious example, when poverty (on whatever scale we are using) reaches 1.40, the probability of a student dropping out is about .80. conversely, at −1.40, the probability is about .20. Beyond these points, the slopes flatten out, giving less change in probability despite rather large changes in *X*.

Think about this relationship in another way. Let's imagine that we were looking at the dosage of a hypothetical drug and the probability

Figure 2.2 The Logistic Curve



Data Source: NELS88, NCES, U.S. Department of Education (http://nces.ed.gov/surveys/nels88/).

that we could cure a disease. The drug is very effective and has no known side effects. If *X* is dosage and *Y* is the probability of cure, you might well get a similar curve. At very low doses, there are very small probabilities of cure, but as the doctors increase the dosage, there will come a point where it begins becoming effective, and as dosage increases (to a point), probability of cure will also increase. Then at some point, the benefit of increasing the dosage will level off, for any number of hypothetical reasons.

This curve, called a logistic curve (among other names), represents the conceptual underpinning of logistic regression—that the probability that something will happen can be predicted from other variables, that these relationships are usually curvilinear, and often asymptotic. There are also often ranges in the observed predictors where much of the "effect" occurs and beyond which increasingly less change is noted.

The Benefits of Odds Over Simple OLS Regression for Binary Outcomes

Returning to our issue of using OLS regression to predict binary outcomes, we can get around the issue of impossible values in OLS regression if we look at odds rather than probabilities. There are drawbacks to odds—such as being difficult for average people to interpret—but they are not constrained by the 0–1 range as probabilities are. Theoretically, odds can range from 0.00 to infinity. Odds are similar to probabilities with an important distinction: the odds of something happening (e.g., dropout) is the probability of that event divided by the probability of the event not happening:

$$\text{Odds}_{(\text{dropout})} = \text{probability of dropout/probability of not dropping out.}$$

$$\text{Odds}_{(\text{dropout})} = \pi/(1-\pi)$$

The odds of a student from a nonpoor family dropping out are about 0.03, and the odds of a student from a poor family dropping out are 0.17. Odds tend to magnify effects (some argue overestimate them), particularly as things get extreme. Although they are mathematically derived from the same data, they can give the uninformed reader the impression of larger effect sizes than is warranted because most readers interpret them as probabilities rather than odds.

The one drawback of OLS regression with this type of data remains once converted to odds—the predicted conditional odds below 0.00 can still be possible. So we have eliminated one issue—predicted odds over 1.0 are legitimate, but the problem of negative odds still remains. So the solution mathematicians and statisticians have come to is to take the natural logarithm of the odds, which has the benefit of having no restriction on minimum or maximum values. Before we talk about *logits,* the natural logarithm of the odds, let's pause and talk about one of the most commonly reported statistics from logistic regression, the odds ratio.

The Odds Ratio

I have heretofore been talking about conditional odds, and now I am going to introduce the term "odds ratio"—which is one of the standard metrics researchers use in logistic regression to understand effects.[4] Conditional

[4]Although I will encourage you to not use it as much as relative risk or predicted probabilities.

odds are the odds that an outcome (i.e., dropping out) will happen given a particular value of another variable (i.e., being below or above average in family income). As you can see in Table 2.1, those are interesting, but without something to compare it to, interpretation is difficult. So the odds ratio has become a standard metric reported in logistic regression. In practice, the odds ratio is the odds of the outcome at one level of X (e.g., 1) relative to the odds of the outcome at another level of X (e.g., 0). In our example in Table 2.1, we only have two levels of X: poor or not poor (1 or 0). If we calculate the ratio of those two odds, we get an odds ratio of 5.67 (0.17/0.03). The interpretation is straightforward (although we will see in the next chapter there are many ways to misinterpret odds ratios, so we will talk that over in detail). In this example, the odds of students from "poor" households dropping out are 5.67 that of students from "not poor" households. This is not a surprising statistic, given what we know of the importance of poverty in predicting educational outcomes, but it is valuable.

In general, odds ratios are calculated as the change in odds for every 1.0 increase in the IV. In the case of binary IVs, it is the comparison of those in the "1" group to those in the "0" group, as you just saw in our example in the previous paragraph. In the case of a continuous IV, the odds ratio would be the change in odds for each increase of 1.0 in the variable of interest. We will talk more about interpretation in coming chapters, but for now let's think of odds ratios as a slope—the change in the conditional odds as the IV increases one unit. It is analogous then to slope, or regression coefficient, in OLS regression. The change in one variable as another variable changes 1.0 in whatever metric the variable is measured. I think of the odds ratio as an analogue to the standardized regression coefficient (beta), as it is metric free, and generally comparable across analyses as an effect size as long as the variables are on a standard metric.

The Logit

The natural logarithm of the odds is called the *logit,* a term that might remind you of the topic of our book—*logistic* regression. And that is no coincidence. We are almost done with our exploration of some of the basics underlying logistic regression, and in this section we come to the crux of the issue—what is the thing that logistic regression is really predicting?

For those of you who are more than a few years removed from high school mathematics, let's do a brief review of a logarithm before continuing. A logarithm is actually a class of mathematical operations where numbers we are used to (which are commonly represented in something called

Table 2.3 Examples of Logarithms

Base:	1,000,000	10,000	100	1	0.01	0.0001	0.000001
2	19.93	13.28	6.64	0	–6.64	–13.28	–19.93
e *(natural log)*	**13.81**	**9.21**	**4.60**	**0**	**–4.60**	**–9.21**	**–13.81**
3	12.58	8.38	4.19	0	–4.19	–8.38	–12.58
4	9.97	6.64	3.32	0	–3.32	–6.64	–9.97
5	8.58	5.72	2.86	0	–2.86	–5.72	–8.58
10	6.00	4.00	1.00	0	–1.00	–4.00	–6.00

"base 10") can be represented by other bases. In brief, a logarithm is the power (exponent) a base number must be raised to in order to get the original number. Any given number can be expressed as Y to the X power in an infinite number of ways. For example, if we were talking about base 10, 1 is 10^0, 100 is 10^2, 16 is $10^{1.2}$, and so on. Thus, $\log_{10}(100) = 2$ ($100 = 10^2$) and $\log_{10}(16) = 1.2$ ($16 = 10^{1.2}$). However, base 10 is not the only option for logarithms—you can literally use any number, although base 10 is one of the more common. Another common option is the natural logarithm, where the constant e (2.7182818 . . .) is the base.[5] In this case, the natural log of 100 is 4.605 ($100 = e^{4.605}$).

As you can see in Table 2.3, the same number can be represented in a variety of ways across a variety of bases. Perhaps more germane to this discussion is the natural logarithm, of base e. If you notice, the natural logarithm of numbers above 1.0 grows from 0 toward infinity as the numbers being log transformed get larger. Interestingly, as numbers go from 1 toward 0, the log of those numbers moves toward infinity in the negative direction.

You may also notice an interesting pattern in these numbers—the log of 100 and the log of 0.01 are identical except for the sign, as are the logs of 10,000 and .0001, and 1,000,000 and 0.000001. This is because in

[5]Sometimes referred to as Euler's number, but usually credited to Bernoulli, who attempted to solve the following formula that was applied to calculations of compound interest: $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$, e has applications in many fields beyond economics and statistics, including being particularly useful in calculus and probability theory, physical sciences, and beyond. It has been calculated to a trillion digits thus far, and like pi, is an enigmatic and interesting number.

exponents, raising something to a negative power (X^{-1}) merely means to calculate 1/X. Thus, the interesting property of logs is that they "pivot" at 1.0, are essentially symmetrical around 1.0, and the log of 100 and 1/100 being identical except for the sign. This is an important revelation that will help with interpreting logistic regression output as you move forward. Trust me on this, and hold on while we march forward!

SUMMARIZING SO FAR ♦

We started off with a hand calculation of simple probabilities and simple odds, and moved into the shortcomings of OLS regression in predicting dichotomous variables—aside from violations of all sorts of assumptions (usually), you can get predicted conditional probabilities (outside the 0 to 1 acceptable range) and conditional odds that are impossible (below 0.00). To handle these shortcomings, the natural logarithm of the odds can conceivably range from $-\infty$ to ∞. Thus, if we use the logit (natural logarithm of the odds) as our dependent variable, we no longer face the issues that probabilities or odds have given us. The dependent variable then becomes logit($\hat{Y}$), and we say that the regression uses a logit link function:

$$\text{Logit} = \textit{natural log}\left(\frac{\pi}{1-\pi}\right)$$

Thus, the regression equation we calculate is:

$$\text{Logit}(\hat{Y}) = b_0 + b_1X_1 \ldots b_kX_k$$

Where b_0 is the intercept, and b_1 is the slope, or the effect of X_1. The equation should look familiar to you if you have ever explored OLS regression—a variable is predicted from a slope and intercept (estimate of Y when $X = 0$). The only difference between this equation and the simple OLS regression equation is the *logit link*. You can think of the concept of a link in regression as a translator of sorts. In the family of regression models, the right-hand side remains relatively similar and what changes is the *link* between the dependent variable and that linear equation. In the case of logistic regression, we use the *logit link* between the dependent variable and the linear regression equation. In other words, we are converting or translating the dependent variable from its natural form, Y, to something that works more effectively, the logit(Y). There are, theoretically, an infinite number of links in regression but only

a few are commonly used in the fields I am familiar with—the linear link or identity link in OLS regression (which essentially means nothing is done to the dependent variable), the logit link (where the dependent variable is converted to logits), and the probit link (which I will discuss in Chapter 9).

So this *logit link* takes care of the conceptual issues of predicting impossible values, but it does not take care of the issue that these types of analyses routinely violate assumptions of OLS regression (or the fact that the logistic curve is nonlinear, meaning that there is no ordinary least squares estimation that will fit data curvilinearly). Thus, logistic regression uses a maximum likelihood estimation (which was discussed briefly in the preceding chapter), which does not rely upon assumptions that are going to be routinely violated when performing this sort of analysis.

It is important to note, however, that when these assumptions of OLS regression are met, OLS regression and logistic regression using maximum likelihood estimation will produce virtually identical coefficients. Yet when these assumptions are not met (specifically, the binary dependent variable violates the assumption of interval or ratio measurement, there is usually heteroscedasticity of residuals, and the assumption of a linear relationship is questionable), which is usually the case with a binary dependent variable, OLS regression will produce inappropriate and/or misestimated results. All this is to say that there is a good reason to use logistic regression when you have a binary outcome—that you will get optimal results for those types of data.

♦ STILL MORE FUN WITH LOGITS, ODDS, AND PROBABILITIES

So this thing that we are going to see in logistic regression is the natural logarithm of the odds of something happening (whatever is 1 when the dependent variable is coded 0 and 1). As you can see from the table above, the log of a thing is difficult for most people who are not professional mathematicians to comprehend in a deep way (or in an accurate way). So in logistic regression, you are going to get the intercept and coefficients in a logit metric (because we are using a logic link function). But most statistical packages also give you "odds ratios" (sometimes abbreviated "OR," or in SPSS, labeled "Exp(B)" because the odds ratio is the exponentiated logit) to make interpretation a bit more straightforward. In the first part of this chapter, we worked our way conceptually and mathematically from numbers to probabilities to odds to logits. It is important to recognize that

these are all essentially the same bit of information, presented in slightly different form. If you have followed to this point, you can see each is a simple mathematical transformation of the other. Because of this, it is also relatively simple to reverse the process. We can start with logits (again, the natural log of the odds of an outcome) and work our way back to conditional probabilities, which are generally easier for people to understand. This is particularly true for those of you who will be communicating to nontechnical audiences (practitioners, policymakers, or the public) and is even useful when talking to other researchers who may not be as well-versed in logistic regression as you are (or will be!).

The Logit Revisited

Most statistical programs will convert your logits to odds ratios for you, but it is valuable to understand the magic that is happening in the software. Below is a sample of the output from SPSS for this same data.

From your knowledge of probability, odds, odds ratios, and logits (natural logarithm of the odds of dropout), much of this might at this point make sense. Let's start with the constant (intercept) first. Remember I have already mentioned that SPSS calls the odds ratio "Exp(B)." In a moment it should become clear why the odds ratio is sometimes called this. For now, know it is the odds ratio. The constant is the predicted logit when $X = 0$ (when students are *not* coming from poor households). Does the 0.03 in the odds ratio (Exp[B] column) look familiar? It should. In Table 2.1, we calculated the odds of dropping out for students from non-poor households (when POOR = 0) to be 0.03. This is the same number from the same data. Recalling that logistic regression doesn't analyze odds or odds ratios, but rather logits, let's look at the left-hand column labeled "B" (regression coefficients)—the constant or intercept is −3.51. That

Table 2.4 SPSS Logistic Regression Output for POOR and DROPOUT Analysis

		B	*SE*	Wald	*df*	Sig.	Exp(B)
Step 1	poor	1.742	.073	566.339	1	.000	5.711
	Constant	−3.514	.066	2793.147	1	.000	.030

Data Source: NELS88, NCES, U.S. Department of Education (http://nces.ed.gov/surveys/nels88/).

won't look familiar yet, but it will in a moment. Recall that we moved from odds to logits by taking the natural log of the odds. In this case, you can calculate the natural log of 0.03, which turns out to be −3.51, which is what we see in this column. In other words, this is the natural log of the odds of dropping out if you are in the "0" category on the independent variable.

Now let's look at the variable of interest, POOR. The odds ratio is 5.71—which is within rounding error of what we calculated by hand, and represents the *change in odds* from POOR = 0 to POOR = 1 (the slope, expressed as an odds ratio). So far so good. And converting to logits, the natural log of 5.71 is 1.742, which is what we see under the regression coefficient column. The rest of the columns involve the tests of significance (whether the odds ratio is significantly different from 1.00, or the logit is significantly different from 0). For example, the Wald statistic is simply the square of the regression coefficient divided by the standard error of the regression coefficient:

$$\text{Wald} = \left(\frac{b}{SE_b}\right)^2 \qquad \text{Eq. 2.1.}$$

I assume you are familiar with significance testing. (You're also probably familiar with the controversy over whether it is even a useful process—I will leave that for others to discuss as we have important substantive issues to confront. We will take null hypothesis statistical testing as the standard practice of the discipline even if it has flaws, and keep moving forward.) If you have a calculator that can handle natural logs, exponents, and such (or access to Excel or similar spreadsheet programs), I encourage you to play with the output from your statistical software like this to help cement your understanding of the relationships between the numbers you are seeing on your output.

♦ WHERE IS THE LOGIT OF THE OTHER GROUP?

So to quickly review: The logit of the intercept is the natural log of 0.03, which equals −3.514. The mathematics of this are relatively straightforward. And the logit of the odds ratio (5.71) is 1.74. To get the natural log of a number, we raise *e* to a particular power.

$$\text{Natural log of } 5.71 = e^{1.74} \qquad \text{Eq. 2.2.}$$

And thus we say the natural log of 5.711 is 1.74. To reverse this, moving from logit to odds, we *exponentiate* the logit—in other words to convert from logit to odds, we raise *e* to the logit power: [6]

$$e^{1.74} = 5.71$$

So the logit of the Poor = 0 group is −3.514, and the exponentiation of that is 0.03. But if you look at Table 2.1 and try to find the logit for the other group (POOR = 1), it is not there, although there is an effect of poor. This is actually a little confusing (at least it was for me when I first tried to figure this all out . . .) until you think carefully about what those numbers are for POOR—they are the effect of a 1.0 increment in POOR—the slope, not the value for that group. The odds ratio is the ratio of two odds (the odds of dropping out if POOR = 1 divided by the odds of dropping out if POOR = 0), as I have already mentioned, the slope or change in odds as you move an increment of 1.0 on the independent variable from POOR = 0 to POOR = 1. This is the same concept we can find in any regression analysis, the common definition of slope. Likewise, the logit of this odds ratio is 1.742. This is the slope, or change in logits between POOR = 0 and POOR = 1. The logit and conditional odds can be found by using the regression equation to predict a logit for that group, and then converting it to odds (or probability):

$$\text{Logit}\,(\hat{Y}) = -3.514 + 1.742(\text{POOR})$$

Using this regression equation we can confirm that the predicted logit for POOR = 0 is −3.514, and the predicted logit for POOR = 1 is

$$\text{Logit}_{(\text{POOR}=1)} = -3.514 + 1.742(1) \text{ or } -1.772$$

CONVERTING FROM LOGITS DIRECTLY ♦ BACK TO CONDITIONAL PROBABILITIES

I will spare you the recapitulation of the mathematics that brought us to this point and merely present the formulae for converting directly from logit (predicted logit, usually) to predicted conditional probability.

$$\text{Conditional probability} = \left(\frac{e^{logit(y)}}{1+e^{logit(y)}}\right)$$

[6]Note that there is minute rounding error in all these calculations. If you are using a scientific calculator, Excel, or some similar process, you use the EXP(*X*) command, where *X* is the logit you want to convert back to an odds ratio.

which is the same as the more spreadsheet-friendly

$$\text{Conditional probability} = \left(\frac{\exp(\text{logit}(y))}{1+\exp(\text{logit}(y))}\right)$$

So we just calculated the logit of POOR = 1 to be −1.772. Using the formula above, that converts directly to 0.145, which is the same conditional probability we calculated by hand at the beginning of this chapter.

Some Benefits of Conditional Probabilities

So why go through all these mathematical machinations? We already have what we want to know when we perform a logistic regression—what variables are significant predictors of the outcome and the magnitude of the relationship (as well as direction), right? One problem we will explore in the next chapter is the fact that odds ratios are problematic in that researchers, practitioners, and the lay public often don't intuitively understand odds—although they often think they do (Davies, Crombie, & Tavakoli, 1998; Holcomb Jr, Chaiworapongsa, Luke, & Burgdorf, 2001), and ratios of things that people don't understand are necessarily even more fraught with difficulty. Odds are not intuitive like probabilities are, and the language needed to technically describe an odds ratio is (as you can see) a bit convoluted.

The situation is not helped by authors' tendency to whitewash this important distinction and use probabilistic language when discussing odds ratios. Even sophisticated researchers will summarize odds ratios using language similar to: "boys are 4.91 times more likely to be recommended to remedial reading than girls" or "boys are 4.91 times as likely to be recommended . . ." when technically the odds ratio should be summarized as "the odds of boys being recommended are 4.91 times the odds of girls being recommended," which does not address exactly what it means for odds to be greater in one group than another. Pedhazur (1997, pp. 760–761) takes great pains to highlight this common error, as do other authors (e.g., Cohen, 2000; Davies, et al., 1998; Holcomb Jr. et al., 2001). Holcomb et al. (2001) report that 26% of authors in top-tier medical journals explicitly misinterpreted odds ratios as relative risk ratios, which are ratios of probabilities rather than odds.

In addition, odds ratios tend to exaggerate effects, as Davies et al. (1998) demonstrate. This effect is particularly egregious when the outcome being examined is not rare (e.g., occurs in more than 5% of the population)

and becomes magnified as the relative risk moves away from 1.0—commonly exceeding 80%–90% inflation in my own simulations.

For this reason, some argue that reporting relative risk (conditional probability of dropout for students from poor households divided by the conditional probability of dropout for students from nonpoor households rather than the conditional odds of one group divided by the conditional odds of another group) is a better practice and is less likely to have biased estimation of effects. As an example, the relative risk in our example is:

$$\text{Relative risk} = 0.145/0.03 = 4.833$$

Comparing our odds ratio to relative risk, we see about a 17.2% inflation with the odds ratio, and as we see in the next chapter, this inflation gets stronger the farther from 1.00 the odds ratio gets. The other advantage is that with relative risk we can use the words "probability" and "likelihood"—words that many of us want to use to describe probabilistic relationships but really should not when discussing odds. The important point is that it is much easier to calculate relative risk and report it when you have all your statistical output in front of you than attempting to piece it together from someone's article, hoping there is enough information.

It is probably more effective to calculate relative risk directly, or, if that is not possible, calculate it from the following formula (Zhang & Yu, 1998):

$$RR = OR / [(1 - P_0) + (P_0 - OR)] \qquad \text{Eq. 2.3.}$$

where RR = relative risk, OR = calculated odds ratio, and P_0 = the proportion of nonexposed individuals (e.g., those lacking the independent variable or the reference group) that experience the outcome in question. This formula is valuable (I have actually used it in some of my publications!), but now that we are logistic regression ninja warriors, slicing and dicing probabilities, odds, odds ratios, and logits, we do not need those sorts of formulae. If you have standard logistic regression output presented in any journal article (an intercept and a slope), you should be able to directly calculate conditional probabilities for each group and relative risks. For example, knowing nothing except the regression equation for this analysis ($\text{logit}_{(y)} = -3.514 + 1.742(\text{POOR})$), we could calculate the predicted probability for POOR = 0 (0.03) and POOR = 1 (0.145) and their ratio (4.833; note that this ratio is a probability ratio, not an odds ratio, which is larger as odds tend to inflate effect sizes).

Table 2.5 Cross-Tabulation of Family Income and Dropout in a More Expanded Form

		dropout		Total	Conditional probabilities	Odds	Predicted logit[1]	Odds ratio	Slope[2]
		.00	1.00						
poor	Yes (1)	7312	1244	8556	0.145	0.170	−1.774	5.67	1.74
	No (0)	7821	233	8054	0.029	0.030	−3.511		
Total		15133	1477	16610					

1. Natural log (probability$_{(group)}$ /1−probability$_{(group)}$)

2. Change in predicted logit as *X* moves from 0 to 1 or logit(group 1)—logit(group 0)

Data Source: NELS88, NCES, U.S. Department of Education (http://nces.ed.gov/surveys/nels88/).

Table 2.6 SPSS Logistic Regression Output for POOR and DROPOUT Analysis

		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1	poor	1.742	.073	566.339	1	.000	5.711	4.947	6.592
	Constant	−3.514	.066	2793.147	1	.000	.030		

Data Source: NELS88, NCES, U.S. Department of Education (http://nces.ed.gov/surveys/nels88/).

Table 2.7 SAS Logistic Regression Output for POOR and DROPOUT Analysis

	Analysis of Maximum Likelihood Estimates				
Parameter	*df*	Estimate	*SE*	Wald Chi-Square	Pr > ChiSq
Intercept	1	−3.5135	0.0665	2793.1509	<.0001
poor	1	1.7423	0.0732	566.3385	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
poor	5.711	4.947	6.592

Data Source: NELS88, NCES, U.S. Department of Education (http://nces.ed.gov/surveys/nels88/).

CONFIDENCE INTERVALS FOR LOGITS, ODDS RATIOS, AND PREDICTED PROBABILITIES

Confidence intervals for point estimates (means, correlations, regression coefficients, intercepts, logits, odds ratios, and probabilities, among others) are desirable because we rarely know for sure the exact population parameter (and because more journals and reviewers want to see them). In the case of simple means, we can estimate the population average from our sample, but our sample is rarely a perfect predictor of the population average. Thus we often calculate 95% confidence intervals from the mean and the standard error, which tells us the range that we are 95 confident the true population mean falls into. Similarly, with regression we can estimate intercepts and slopes, but again we need to calculate confidence intervals to honestly communicate the range the true population parameter might fall into. Most statistical programs will provide these for you, but they are not difficult to calculate:

$$\text{Confidence interval} = \text{statistic} \pm 1.96(SE_{\text{statistic}})$$

Thus, taking the logistic regression from our example in this chapter, SPSS tells us that the odds ratio is 5.711, with a 95%CI of 4.95, 6.59. This means that we are 95% confident that the real population odds ratio falls between 4.95 and 6.59. That is a relatively large range. Similarly, we could calculate the 95%CI for the logit of POOR as

$$95\%\text{CI (POOR)} = 1.742 \pm 1.96(0.073)$$

This turns out to be 1.599, 1.885. Knowing what we know about converting between logits and odds ratios, we can easily exponentiate these

two numbers to check that we get the same odds ratios that SPSS served up for us. In fact, when we do this, we get 4.95 and 6.59, as expected. Perhaps more interestingly, we can also compute 95%CIs for predicted probabilities. Recall that we converted these two logits for POOR = 0 and 1 (–3.514 for POOR = 0 and –1.774 for POOR = 1) to conditional probabilities of 0.03 and 0.145. But those are only point estimates, and we might need to fit confidence intervals around them. We can produce confidence intervals around those predicted probabilities after calculating the standard error of the predicted probability:

$$SE\ \hat{\pi} = \hat{\pi}\ (1 - \hat{\pi})\ *\ SE_{\text{logit}}$$

$$SE\ \hat{\pi} = 0.145(1–0.145)\ *\ 0.073 = 0.00905$$

And then using the standard formula for 95% confidence intervals for POOR = 1:

$$\text{Upper 95\%CI} = 0.145 + 1.96*0.00905 = 0.163$$

$$\text{Lower 95\%CI} = 0.145 - 1.96*0.00905 = 0.127$$

We can also calculate these confidence intervals for predicted probability directly from statistical output if you have the 95%CIs for predicted logit or predicted odds ratio. As mentioned earlier, it is simple to calculate a 95%CI for the logit of this group:

$$= (–1.772) \pm 1.96*SE_{\text{logit}}$$

which also ends up being –1.629 and –1.915. Converting these logits to probabilities, we get a 95%CI in probabilities of 0.128 and 0.164, both of which are reasonably close to the 0.127 and 0.163 we calculated from the predicted probability and the $SE\ \hat{\pi}$, above.

You might find all this converting and cross-converting a bit boring or bothersome or redundant. What I find comforting and interesting about it is that it reinforces confidence that you know how these numbers work and what they mean. And by successfully manipulating them, getting to the identical conclusions via multiple paths, we are confident that the math is right. This leaves you in a powerful position to communicate the results of logistic regression to your audience in the best, clearest manner possible.

ONE CONCRETE REASON WHY YOU SHOULD CARE ♦ ABOUT THIS STUFF—CLARITY OF COMMUNICATION

One of the reasons I started investigating all this "under the hood" stuff—what are we really calculating when we do logistic regression and what all those numbers mean—is because of my desire to be able to clearly communicate my findings to technical and nontechnical audiences. It is all well and good to report odds ratios (or relative risk) and their confidence intervals for simple effects like we have in these examples. It is relatively easy for people to understand that the odds of students in poor households dropping out of high school are 5.71 that of students not living in poor households. But what if you are not reporting a simple effect, but rather a curvilinear effect, or an interaction? I always try to graph those out, and when I present them in publications or at conferences, people seem a bit perplexed at what is being graphed. Concretely, if you graph out a logistic regression analysis, you are graphing in units of logits. But even those of us who use logistic regression routinely have trouble conceptualizing logits. Most researchers and consumers of research would be more comfortable talking about odds or probabilities. And now that you know how simple it is to convert from logits or odds ratios to predicted probabilities, there is no reason not to make your graphs and tables as easily comprehensible for the reader as possible.

Later in the book we will get into graphing curvilinear effects and interactions (which I think are almost always more interesting than simple effects). By understanding everything in this chapter, you will be well prepared to accurately and appropriately communicate more complex results. I will show you how to graph out curvilinear effects and interactions, and I will also show you that graphing in logits is not only confusing to readers, but also can be very misleading. So we will practice graphing the results of logistic regressions in conditional probabilities

ENRICHMENT

1. Calculate the following:
   a. Convert the following odds ratios to logits:
      i. 1.00
      ii. 3.00
      iii. 0.333333

iv. 1.50
v. 0.50
vi. Note that the natural logs of 3.00 and 0.33333 are "symmetrical," or the same number just with a negative sign (within rounding). Why are the logits for 1.50 and 0.50 not symmetrical as 3.00 and 0.33333 is?

b. Convert the following logits to conditional probabilities:
i. 0.00
ii. 0.50
iii. 1.00
iv. 2.00
v. −0.50
vi. −1.00
vii. −2.00

2. On my website for this book, under Chapter 2, there is a sample of 3328 students from the 2002 Education Longitudinal Study (ELS2002; http://nces.ed.gov/surveys/els2002/) from NCES. With help from one of my wonderful students (LaShauna Dean-Nganga), we compiled this small data set so that you can explore the same sort of calculations that I went through in this chapter. In this new example for you to work through, we are looking at student postsecondary plans (0 = not planning on attending a 4-year postsecondary institution following graduation from high school, 1 = planning on attending a 4-year postsecondary institution following high school) and a variable I am calling RICH (0 = student family income < $35,000.00 per year in 2002, 1 = student family income above $35,000.00 per year in 2002).[7]

a. Download the data file from the book website.

b. Produce a cross-tabulation of BACHPLAN versus RICH as I did in Table 2.1 (a template is provided below for your convenience to guide your calculations).

c. Compute conditional probabilities for RICH and not RICH students.

d. Compute conditional odds for RICH and not RICH students. You should see that those coming from more affluent homes are more likely to plan on attaining their 4-year degree (and you should not be surprised by this!).

e. Compute the odds ratio.

f. In SPSS, SAS, R, or a statistical computing package of your choosing, perform the same analysis. Verify you receive the same (or similar to rounding error) odds ratio.

g. Compute the natural logarithm of the odds ratio. Verify it is the same as the B (or logistic regression coefficient for RICH) produced by your software.

h. Examine the odds ratio for the constant (intercept where $X = 0$). Verify it is identical to the conditional odds of RICH = 0 (within rounding error).

i. Reverse engineer the process. Exponentiate the logit for RICH and verify that you get the odds ratio (in Excel, e.g., the function for exponentiation is = EXP(X), where X is the odds ratio). Using the formula, convert from EXP(b) to probability.

[7]Obviously I call the variable "RICH" not to indicate that a family income in the United States of $35,000.00 a year indicates wealth, but rather for a convenient label. Don't send me e-hate-mail over the label!

Enrichment Table 2.1 Worksheet

		BACHPLAN		Total	Conditional probabilities	Conditional Odds	Logit	Odds ratio	Slope
		0	1						
RICH	Yes (1)								
	No (0)								
Total									

ANSWER KEY

1. a. i. 0.000
ii. 1.099
iii. −1.099

iv. 0.405
v. –0.693
vi. Because 1.50 is only 50% bigger than 1.0, whereas 0.50 is one half the size. 2.0 would be "symmetrical" with 0.50 because they represent a doubling or halving (dividing by 2 or multiplying by 2, just as 3.00 and 0.3333 are a multiplying or dividing by 3). This is a source of common errors and misinterpretations in logistic regression.

b. i. 0.50
ii. 0.622
iii. 0.731
iv. 0.881
v. 0.378
vi. 0.269
vii. 0.119

Enrichment Table 2.2 Cross-Tabulation of RICH and BACHPLAN

		BACHPLAN		Total	Conditional probabilities	Conditional Odds	Logit	Odds ratio	Slope
		0	1						
RICH	Yes (1)	257	1563	1820	.859	6.08	1.805	1.575	0.454
	No (0)	310	1198	1508	.794	3.86	1.351		
Total		567	2761	3328					

Some notes:

The logits are the natural logarithms of the conditional odds. Interestingly, as we will see next chapter, you will not see these two logits in SPSS or SAS output. Rather, you will see the log of the odds for the intercept, and the log of the odds ratio for the slope (b) of the IV. This makes sense, as the b is the slope or effect of the IV, not the log of the odds of the other group. Once we realize that an odds ratio is a slope (the change in odds as you increase 1.0 in the IV) and that the regression coefficient is the natural log of the odds ratio or the change in logits from 0 to 1 (again, the slope), this might make things a bit clearer.

SPSS Logistic Regression Output

Enrichment Table 2.3 Variables in the Equation

		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	RICH	.453	.093	23.933	1	.000	1.574	1.312	1.887
	Constant	1.352	.064	450.055	1	.000	3.865		

a. Variable(s) entered on step 1: RICH.

Alternative Statistical Computing Package Output: SAS

Enrichment Table 2.4

Analysis of Maximum Likelihood Estimates					
Parameter	*df*	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.3517	0.0637	450.0176	<.0001
RICH	1	0.4535	0.0927	23.9346	<.0001

Enrichment Table 2.5

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
RICH	1.574	1.312	1.887

SYNTAX EXAMPLES

Syntax for Performing Analysis in SPSS

```
LOGISTIC REGRESSION VARIABLES RICH
/METHOD = ENTER BACHPLAN
/SAVE = ZRESID
/PRINT = CI(95)
/CRITERIA = PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Syntax for Performing Analysis in SAS

```
PROC LOGISTIC DATA = book.CH2_ex descending;
MODEL bachplan = rich;
output out = results p = predict;
run;
```

REFERENCES

Cohen, M. P. (2000). Note on the odds ratio and the probability ratio. *Journal of Educational and Behavioral Statistics, 25*(2), 249–252.

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal, 316,* 989–991.

Holcomb Jr, W. L., Chaiworapongsa, T., Luke, D. A., & Burgdorf, K. D. (2001). An odd measure of risk: use and misuse of the odds ratio. *Obstetrics & Gynecology, 98*(4), 685.

Osborne, J. W. (2011). Best practices in using large, complex samples: The importance of using appropriate weights and design effect compensation. *Practical Assessment Research & Evaluation, 16*(12), 1–7.

Pedhazur, E. J. (1997). *Multiple Regression in behavioral research: Explanation and prediction.* Fort Worth, TX: Harcourt Brace College Publishers.

Zhang, J., & Yu, K. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association, 280,* 1690–1691.